Teaching Large Language Models to Reason with Reinforcement Learning

Year: 2024 | *Citations: 139* | *Authors: Alex Havrilla, Yuqing Du, S. Raparthy*

Abstract

Reinforcement Learning from Human Feedback (\textbf{RLHF}) has emerged as a dominant approach for aligning LLM outputs with human preferences. Inspired by the success of RLHF, we study the performance of multiple algorithms that learn from feedback (Expert Iteration, Proximal Policy Optimization (\textbf{PPO}), Return-Conditioned RL) on improving LLM reasoning capabilities. We investigate both sparse and dense rewards provided to the LLM both heuristically and via a learned reward model. We additionally start from multiple model sizes and initializations both with and without supervised fine-tuning (\textbf{SFT}) data. Overall, we find all algorithms perform comparably, with Expert Iteration performing best in most cases. Surprisingly, we find the sample complexity of Expert Iteration is similar to that of PPO, requiring at most on the order of 10^6 samples to converge from a pretrained checkpoint. We investigate why this is the case, concluding that during RL training models fail to explore significantly beyond solutions already produced by SFT models. Additionally, we discuss a trade off between maj@1 and pass@96 metric performance during SFT training and how conversely RL training improves both simultaneously. We then conclude by discussing the implications of our findings for RLHF and the future role of RL in LLM fine-tuning.